Deep Learning Techniques in the Textual Data Mining of Chinese Medicine Literature
Evaluating AI-Generated Responses on Cancer Pathogenesis and Treatment

Introduction

Objectives
This study evaluates an AI model (DeepSeek-R1 7B) trained on TCM literature to generate responses on cancer pathogenesis and treatment. We assessed the model's accuracy, clarity, and clinical applicability, focusing on lung and breast cancer.

Significance
- Addresses gaps in scientific validation of TCM mechanisms (e.g., acupuncture, herbal medicine).
- Provides a knowledge base for clinicians/researchers via AI-driven insights.

Key Focus Areas: Lung cancer & breast cancer pathogenesis, integration of TCM theory with modern oncology.

Results

❖ Accuracy & Relevance: Trained responses scored higher (avg. 4.0–5.0) in scientific validity and direct relevance.
❖ Comprehensibility: Individuals without TCM background found jargon challenging (scores: 3.0–4.0); trained responses were clearer but needed simplification.
❖ Depth of details: Trained responses provided deeper explanations (avg. 4.0–5.0) but sometimes at the cost of comprehensibility.
❖ Clinical Utility: Trained outputs were rated more actionable (avg. 3.0–5.0), especially for TCM-WM integration.

Participant Feedback
- 60% noted consistent differences (e.g., one response set was more detailed).
- 62.5% TCM student-respondent rated terminology use as "partially correct."
- Suggested improvements: Simpler language, more scientific references, and better TCM-WM integration.
- Limitations: Occasional over-complexity in responses.

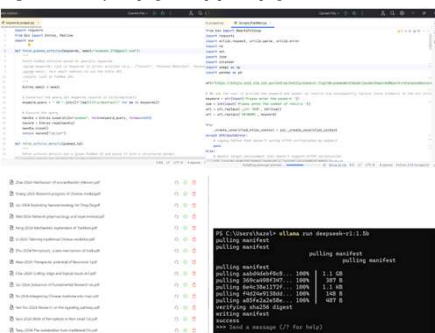**Figure 1a & 1b.** Python program for paper scraping

Figure 3. Interface for database in Cherry Studio

Figure 4. Running initial local deployment (without using interface)

Methods and Materials

1. **Data Collection & Curation**
   ➤ Tools: Python crawler, EndNote
   ➤ Process:
      - Scraped 8,000+ papers from PubMed/CNKI → filtered to 1000+ high-relevance studies.
      - Keywords: "TCM," "cancer," "mechanism," "acupuncture," "lung/breast cancer."
   ➤ Challenge: Irrelevant results → resolved via elaboration in prompt engineering

2. **Model Deployment**
   ➤ Tools: DeepSeek-R1-7B (local), Cherry Studio (interface)
   ➤ Process:
      - Local deployment in terminal
      - Switched from failed WebUI and Chatbox interface to Cherry Studio for stable API integration.

3. **Model Training**
   ➤ Approach: prompt engineering
      - Designed prompts with constraints (e.g., "Cite sources from uploaded papers").
   ➤ Blind testing using 3 questions on 'cancer', 'lung cancer' and 'breast cancer' respectively.
   ➤ User Feedback: Tested responses with TCM students and the public; improvements include:
      - Simplified language (e.g., analogies like "Acupuncture acts as a traffic controller for energy flow").
      - Structured layouts (bullets, headers).
      - Graphical representations (e.g. synergistic outcomes of TCM and WM treatments)

Table 1. Comparison across Trained models (T1 & T2) and Control

Conclusions

The AI model shows promise in bridging TCM and modern oncology but requires refinement in clarity and terminology.

Next steps:
1. Expand datasets with non-English literature.
2. Validate outputs in clinical settings.
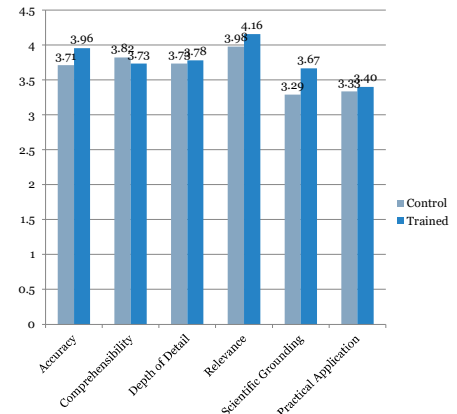3. Enhance interdisciplinary training for seamless TCM-WM integration.

Chart 1. Comparison of scores of Control vs Trained model.

Discussion

This project demonstrates the potential of AI in synthesizing TCM and modern oncology knowledge. The trained DeepSeek-R1 model showed measurable improvements in scientific accuracy, depth of mechanistic detail, and relevance compared to its untrained counterpart, validating the effectiveness of domain-specific fine-tuning. Responses enriched with TCM-WM integration, such as linking Qi deficiency to immunosuppressive tumor microenvironments or explaining herb-chemotherapy interactions, were rated highly by both TCM students and the public. This suggests AI can serve as a bridging tool for interdisciplinary dialogue in integrative oncology.

Figure 2. Sample output of control (left) and trained model (right)

References
1. Pan, D., Guo, Y., Fan, Y., & Wan, H. (2024). Development and Application of Traditional Chinese Medicine Using AI Machine Learning and Deep Learning Strategies. The American Journal of Chinese Medicine (1979), 52(3), 605–623. https://doi.org/10.1142/S0192415X24500265
2. Li, D., Hu, J., Zhang, L., Li, L., Yin, Q., Shi, J., Guo, H., Zhang, Y., & Zhuang, P. (2022). Deep learning and machine intelligence: New computational modeling techniques for discovery of the combination rules and pharmacodynamic characteristics of Traditional Chinese Medicine. European Journal of Pharmacology, 933, 175260–175260. https://doi.org/10.1016/j.ejphar.2022.175260
3. Li, R., Ren, G., Yan, J., Zou, B., & Liu, Q. (2024). Intelligent question answering system for traditional Chinese medicine based on BSG deep learning model: taking prescription and Chinese materia medica as examples. Digital Chinese Medicine (Online), 7(1), 47–55. https://doi.org/10.1016/j.dcmed.2024.04.006